Modélisation stochastique et spectrale de l'occupation du sol

Jean François Mari, Odile Horn

▶ **To cite this version:**

Jean François Mari, Odile Horn. Modélisation stochastique et spectrale de l'occupation du sol. SFC 2019, Sep 2019, Nancy, France. hal-02429701

HAL Id: hal-02429701
https://inria.hal.science/hal-02429701

Submitted on 6 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modélisation stochastique et spectrale de l'occupation du sol

Jean François Mari*, Odile Horn**

*Université de Lorraine, Loria, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France
jfmari@loria.fr
** LCOMS,ISEA, Université de Lorraine, 7, rue Marconi, Metz, F- 57070, France
odile.horn@univ-lorraine.fr

Résumé. Nous proposons une approche stochastique pour découvrir des items périodiques dans un processus stationnaire. Le passage d'une représentation sous forme d'une série temporelle d'items à une représentation sous forme d'une série temporelle de tenseurs multi-dimensionnels nous permet d'utiliser les techniques de traitement de signaux multi-dimensionnels. A l'aide des coefficients d'auto corrélation croisés, nous montrons sur des données artificielles qu'une analyse spectrale permet de faire apparaître des comportements périodiques.

1 Introduction

L'activité humaine est source de données temporelles de différentes natures : données numériques telles que les prix de l'essence à la pompe ou données catégorielles telles les majorités parlementaires de la cinquième république.

Dans toutes ces séries de données, l'observation de comportements périodiques présente un intérêt pour l'extraction de connaissances. Nous nous plaçons dans le cas d'une source de données stationnaire, c'est à dire dont la statistique sur une fenêtre temporelle glissante est indépendante du temps et dont les réalisations sont des séries temporelles de catégories appelées items.

En ce qui nous concerne, nous nous intéressons aux occupations annuelles d'une parcelle agricole sous la forme des cultures qui y sont portées. Tant que l'ensemble des opportunités ou contraintes économiques et climatiques imposées aux cultivateurs ne changent pas, on peut faire l'hypothèse que les séries de cultures dans les différentes parcelles se ressembleront tout en acceptant une certaine variabilité inhérente à toute activité humaine.

La recherche de périodes dans des séries temporelles d'items est principalement faite à l'aide de méthodes combinatoires Elfeky et al. (2005); Tatavarty et al. (2007); Galbrun et al. (2018). Toutes utilisent des seuils fixés *a priori* pour décider si la répétition d'un comportement est significatif et périodique.

Les séries de données échantillonnées à partir d'un signal réel continu sont traitées par une analyse spectrale des fonctions d'auto corrélation depuis les travaux de Vlachos et al. (2005) et Li et al. (2012).

L'originalité de notre travail est double : d'une part, nous proposons une méthode de traitement de données catégorielles employant les formalismes de l'analyse spectrale. Cela nécessite

une représentation tensorielle de ces données afin d'utiliser les techniques éprouvées d'analyse des signaux numériques.

D'autre part, pour pallier la trop courte durée des séries temporelles de données disponible, nous traitons plusieurs séries d'items simultanément en les considérant comme des échantillons d'une source probabiliste de séries temporelles dont les moments peuvent être calculés efficacement.

Cet article est structuré de la façon suivante. Après avoir fixé le cadre de la modélisation stochastique d'un champ de tenseurs, nous proposons d'utiliser les coefficients d'auto corrélation croisée entre dimensions des tenseurs pour faire apparaître par analyse spectrale des comportements périodiques. Nous décrivons ensuite un générateur de séquences périodiques à l'aide de modèles de Markov cachés (HMM) et utilisons les données produites pour les analyser et extraire leurs périodes. Enfin, dans une conclusion / discussion, nous esquissons une méthode pour dépasser la recherche de périodes sur un seul item en la généralisant à la détection de motifs périodiques impliquant plusieurs items.

2 Processus Stochastique

Considérons une séquence temporelle $x_1, x_2, \dots x_T$ de T items issus d'un ensemble $\mathcal{E} = \{e_1, e_2, \dots e_K\}$ de K différentes catégories comme la série temporelle des T occupations du sol (LU comme *Land Use*), observées aux instants $1, 2, \dots T$ sur une parcelle agricole dans un territoire donné. Chaque LU appartient à l'ensemble $\mathcal{E}$. Les x_t sont les réalisations d'une variable aléatoire $X_t(\omega)$ aux instants t sur une parcelle agricole représentée par ω.

Supposons avoir enquêté une mosaïque parcellaire pour relever toutes ses LU pendant T années. Ces données définissent une matrice M dans laquelle la ligne i représente les T LU de la parcelle i observées aux années $1, 2, \dots T$. La colonne t représente les LU enquêtées dans tout le territoire à l'instant t. Cette matrice est un échantillon de la série temporelle des variables aléatoires $X_1(\omega), X_2(\omega), \dots X_T(\omega)$.

Afin de pouvoir utiliser les méthodes de traitement du signal, nous représentons ces données comme un champ de tenseurs de $\mathcal{R}^K$. La série temporelle de LU sur une parcelle ω sera représentée par une séquence de T vecteurs appartenant à $\mathcal{R}^K$. Le vecteur δ^i ayant toutes ses composantes à 0 exceptée la i^e égale à 1

$$\delta^i = \begin{bmatrix} 0 \\ \dots \\ 1 \\ \dots \\ 0 \end{bmatrix} \leftarrow i$$

représentera la catégorie e_i. La composante i du vecteur δ^i_t est la fonction Dirac qui représente l'observation de e_i comme fonction de t. Le champ de tenseurs remplace la série temporelle d'items par une série de fonctions de Dirac multidimensionnelles permettant ainsi l'utilisation des techniques associées aux signaux aléatoires multidimensionnels.

2.1 Moment d'ordre deux croisé

Nous définissons le second moment croisé par l'espérance :

$$\begin{aligned} C_X(\tau) &= E\left[X_t(\omega)X^*_{t+\tau}(\omega)\right] \\ &= \frac{1}{T}\sum_t x_t x^*_{t+\tau} Prob(x_t, x_{t+\tau}) \end{aligned} \tag{1}$$

dans lequel x^* représente la transposé du vecteur x.

Chaque terme $x_t x^*_{t+\tau}$ dans la somme de l'équation 1 est une matrice $K \times K$. Quand $(X_t, X_{t+\tau})$ prend les valeurs (δ^i, δ^j) aux instants $(t, t+\tau)$ le produit $\delta^i_t \delta^{j*}_{t+\tau}$ est une matrice nulle avec un 1 à l'indice (i, j).

$$\delta^i_t \delta^{j*}_{t+\tau} = \begin{array}{c} \left[\begin{array}{ccccc} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{array}\right] \leftarrow i \\ \begin{array}{ccccc} & & \uparrow & & \\ & & j & & \end{array} \end{array}$$

Le terme général (i, j) de $C_X(\tau)$ est :

$$C_X(\tau)[i, j] = \frac{1}{T}\sum_t Prob(\delta^i_t, \delta^j_{t+\tau}) \tag{2}$$

δ^i_t signifie qu'à l'instant t l'item e_i a été observé. La probabilité $Prob(\delta^i_t, \delta^j_{t+\tau})$ peut être estimée à l'aide du nombre d'occurrences dans la matrice M du couple (e_i, e_j) dans les colonnes t et $t+\tau$ respectivement.

2.2 La fonction d'auto covariance

La fonction d'auto covariance est définie à partir de l'équation 1 mais avec des variables centrées.

$$\begin{aligned} R_{XX}(\tau) &= E\left[(X_t(\omega) - E\left[X_t(\omega)\right])(X^*_{t+\tau}(\omega) - E\left[X^*_{t+\tau}(\omega)\right])\right] \\ &= E\left[X_t(\omega)X^*_{t+\tau}(\omega)\right] - E\left[X_t(\omega)\right] E\left[X^*_{t+\tau}(\omega)\right] \end{aligned} \tag{3}$$

Dans un processus stationnaire, l'équation 3 devient

$$R_{XX}(\tau) = E\left[X_t(\omega)X^*_{t+\tau}(\omega)\right] - E^2\left[X(\omega)\right] \tag{4}$$

Le terme général (i, j) est égal à :

$$R_{XX}(\tau)[i, j] = \frac{1}{T}\sum_t Prob(\delta^i_t, \delta^j_{t+\tau}) - E^i E^j \tag{5}$$

E^i est la composante i du vecteur $E\left[X(\omega)\right]$.

3 Expérimentation sur des données artificielles

3.1 Génération de données périodiques artificielles

Afin de démontrer l'intérêt de $R_{XX}(\tau)$ pour révéler des items périodiques, nous avons construit des séquences présentant des comportements périodiques à l'aide d'un modèle de Markvov caché (HMM) (*cf.* figure 1).

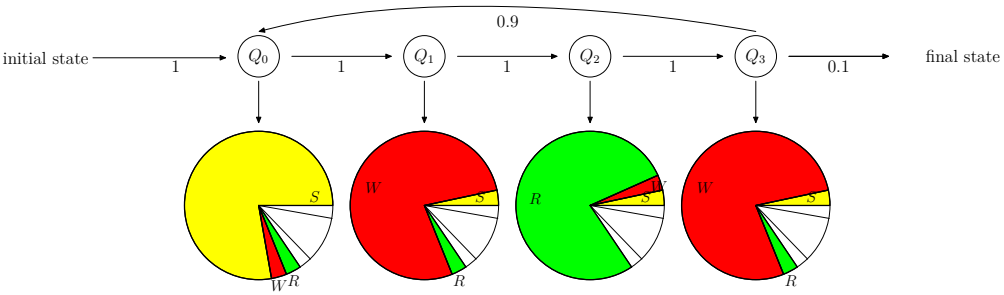


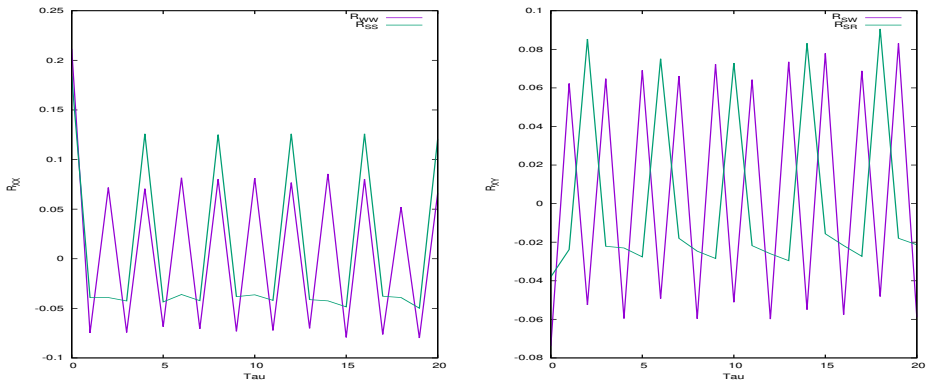
FIG. 1 – *HMM pour simuler des répétitions du patron [S-W-R-W]. Les Q_i sont les états du HMM. Les camemberts de secteurs colorés représentent les probabilités utilisées pour générer les différents items à chaque état. Tous les items ne sont pas représentés*

Nous avons défini tout d'abord un ensemble $\mathcal{E} = \{S, W, R, \ldots\}$ de 11 labels représentant les 11 LU rencontrés dans les parcelles : Tournesol (*Sunflower*), Blé (*Wheat*), Colza (*Rapeseed*),... Nous avons ensuite construit 8 répétitions du patron [S-W-R-W] pour obtenir une séquence de 32 symboles représentant la suite des LU sur une parcelle pendant 32 années. Afin de bruiter cette séquence, des substitutions de symboles ont été effectuées. Ce processus a été répété pour construire 50 séquences de 32 symboles chacune. Le but de cette expérience est de retrouver une période de 4 pour les symboles R et S ainsi qu'une période de 2 pour le symbole W.

Pour simuler ces séquences, le HMM se comporte comme un automate dans lequel un jeton se déplace aléatoirement depuis l'état initial jusqu'à l'état final en fonction des transitions possibles entre états. A chaque état, un label issu de $\mathcal{E}$ est aléatoirement produit à l'aide de la densité de probabilité (pdf) associée à cet état. Le processus est répété jusqu'à produire une séquence de la longueur désirée.

Différentes pdf de différentes entropies sont utilisées pour simuler différents niveaux de bruit dans les séquences. L'entropie est une mesure de l'imprédictibilité d'une pdf. On la mesure en nombre de bits nécessaires pour numéroter et distinguer les évènements. Elle est maximale pour la loi uniforme – tous les évènements doivent être numérotés – et nulle pour une loi certaine : il n'y a rien à numéroter. Pour une pdf possédant n symboles de probabilité $p_i, i = 1, n$, elle est définie par :

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{6}$$



(a) Périodes de S (=4) et W (=2) (b) Décalages (lags) entre S-W (=1) et S-R (=2)

FIG. 2 – *Signal d'auto corrélation calculé sur des séquences générées par le HMM décrit Fig. 1 construit avec des pdf d'entropie 2,65 bits*

3.2 Analyse spectrale des fonctions d'auto corrélation

En traitement du signal, la fonction réelle d'auto corrélation $R_{XX}(\tau)(i,i)$, $\tau = 0, T-1$ est utilisée pour révéler les comportements périodiques à l'aide d'une analyse spectrale donnée par une transformée de Fourier discrète rapide (FFT). Dans le cas particulier de signaux artificiels, l'observation de la courbe $R_{XX}(\tau)(i,i)$ peut suffire pour révéler des comportements périodiques. La période est le petit décalage – ou retard – en temps (*lag* en anglais) où la fonction d'auto corrélation atteint un maximum. La figure 2(a) montre que le symbole "W" (courbe mauve) a une période de 2 – ou un temps de retour de 2 dans le langage des agronomes – , alors que le symbole "S" (courbe verte) a une période de 4. Nous développons aussi une analyse spectrale qui pourra servir à détecter ces mêmes motifs sur des signaux réels bruités. Dans ce sens, nous cherchons une représentation fréquentielle par une transformée de Fourier rapide.

$R_{XX}(\tau)(i,i)$ est tout d'abord fenêtré l'aide d'une fenêtre de Hamming de longueur 32 afin d'atténuer les problèmes liés au caractère limité de la séquence.

Lorsque nous notons $f_i(\tau) = R_{XX}(\tau)(i,i)$ une fonction de τ paramétrée par i, la FFT discrète $\mathcal{F}_i(k)$ sur les $N = 32$ points est définie par :

$$\mathcal{F}_i(k) = \sum_{\tau=0}^{N-1} f_i(\tau) \exp -j\frac{2\pi k\tau}{N}, \quad k = 0, \cdots N-1 \tag{7}$$

Les valeurs $\frac{k}{N}$ sont appelées les fréquences alors que $\frac{N}{k}$ représentent les périodes. La représentation graphique du module $\parallel \mathcal{F}_i(k) \parallel$ vu comme une fonction de k/N est appelée un spectrogramme et vu comme une fonction de N/k un périodogramme. Une valeur de $N = 32$ permet une résolution fréquentielle raisonnable pour le genre de signaux simulés. Leur pendants réels doivent être des résultats d'enquêtes annuelles de terrain. Il est irréaliste d'envisager des séquences de longueur supérieure à 32. La figure 3 montre le spectrogramme $\parallel \mathcal{F}_i(k) \parallel$ pour les items "S" et "R".

Afin de détecter les pics dans un périodogramme, nous avons suivi la stratégie développée par (Li et al., 2012; Vlachos et al., 2005). Chaque série artificielle est réarrangée aléatoirement

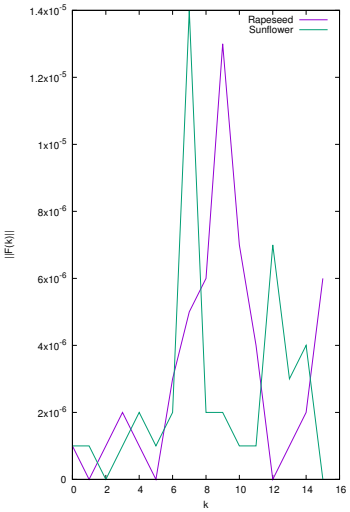


FIG. 3 – *Spectrogramme des items "R" et "S". Les ordonnées représentent* $\| \mathcal{F}_i(k) \|$ *pour les items* $i = Rapeseed$ *et* $i = Tournesol$*. L'axe des abscisses représente le* k *de la fréquence* $\frac{k}{N}$*. Dans une fenêtre de longueur (*$N = 32$*) chaque item montre une fréquence de* $\frac{8}{32}$ *soit une période de* $\frac{32}{8} = 4$*. Les pics de part et d'autre de* $k = 8$ *correspondent à la fréquence 1/4. L'imprécision du résultat provient du faible nombre de valeurs* $N = 32$ *de la série*

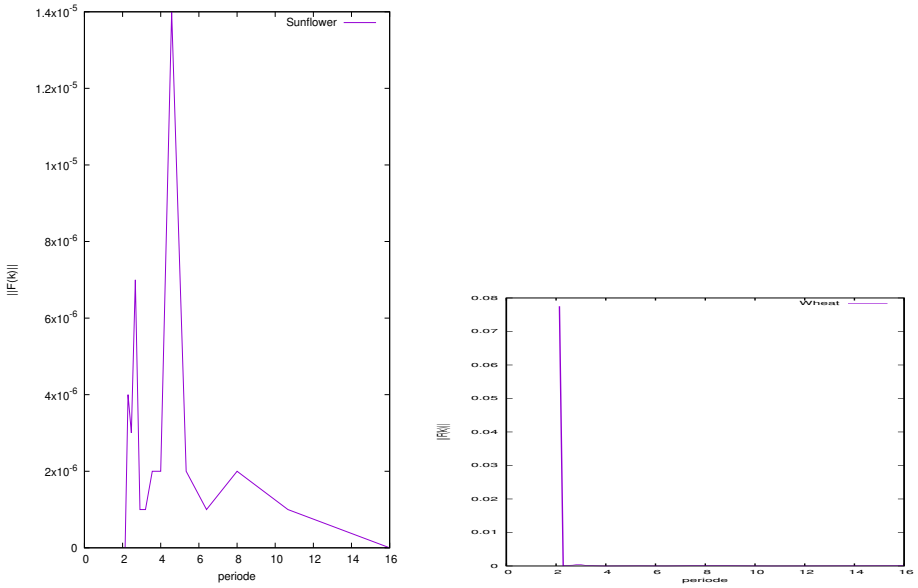
pour faire disparaître tout comportement périodique. Après le calcul de la fonction d'auto corrélation, son fenêtrage et sa FFT, le maximum du périodogramme est stocké. Cent réarrangements sont effectués pour calculer moyenne et écart type des différents maximums. Le seuil qui détermine si un pic est significatif est placé à la moyenne plus deux fois l'écart type ce qui correspond à une confidence de 95%.

$R_{XX}(\tau)(i,j), i \neq j$ donne aussi une indication sur la co-occurrence des symboles e_i et e_j. Fig. 2(b) montre que la corrélation entre les symboles "S" et "W" au décalage (*lag*) de 1 est plus important qu'à 2 ce qui signifie que le couple "S-W" est plus fréquent que le patron "(S- ?-W)" (le caractère joker " ?" représente n'importe quel autre symbole appartenant à $\mathcal{E}$).

Une analyse similaire peut être faite sur le signal $R_{XX}(\tau)(i,j)$ lorsque (i,j) représente les symboles "S" et "R" et montre un maximum au décalage 2 signifiant que "S- ?-R" doit être suspecté. Donc, $R_{XX}(\tau)(i,j)$ peut être utilisé comme un trait pour hypothétiser des successions de symboles dans des séries temporelles bruitées.

4 Conclusion

Cet article décrit une méthode hybride symbolique et numérique pour analyser un ensemble de séries temporelles d'items afin d'en extraire des comportements périodiques. Le passage d'une représentation sous forme d'une série temporelle d'items à une représentation sous forme d'une série temporelle de tenseurs multidimensionnels nous permet d'utiliser les techniques de traitement de signaux multidimensionnels.



(a) Sunflower (b) Wheat

FIG. 4 – *Périodogramme des signaux d'auto corrélation de "S = Sunflower" (période = 4) et "W = Wheat" (période = 2)*

En considérant toutes ces séquences comme un échantillonage d'une série temporelle de variables aléatoires $X_t(\omega)$, la fonction d'auto corrélation peut être calculée efficacement et détecte la périodicité d'un item que l'on peut valider par une analyse spectrale donnée par FFT ainsi que la périodicité de la co occurrence de deux items à plus ou moins longue distance. Ces propriétés sont des indices qu'un analyste utilise pour retrouver des séquences périodiques.

Ce travail se poursuivra en traitant des signaux réels issus de processus vivants caractérisés par une grande variabilité – l'occupation de parcelles agricoles dans une région donnée – et en évaluant l'apport de la méthode numérique développée dans cet article aux méthodes combinatoires. Dans le cadre d'une fouille de données, une interaction serrée avec un agronome permettra de détecter et quantifier des rotations entre cultures dans une région agricole. Nous envisageons ainsi la spécification d'un algorithme de détection de rotations de culture qui pourrait constituer une alternative à l'expertise d'un analyste agronome Le Ber et al. (2006); Schaller et al. (2012).

Références

Elfeky, M. G., W. G. Aref, et A. K. Elmagarmid (2005). Periodicity detection in time series databases. *IEEE Transactions on Knowledge and Data Engineering 17*(7), 875–887.

Galbrun, E., P. Cellier, N. Tatti, A. Termier, et B. Crémilleux (2018). Mining Periodic Patterns with a MDL Criterion. In *ECML/PKDD 2018 European Conference on Machine Learning*

and Principles and Practice of Knowledge Discovery in Databases, Dublin, Ireland, pp. 535–551.

Le Ber, F., M. Benoît, C. Schott, J.-F. Mari, et C. Mignolet (2006). Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software. *Ecological Modelling 191*(1), 170 – 185.

Li, Z., J. Han, B. Ding, et R. Kays (2012). Mining periodic behaviors of object movements for animal and biological sustainability studies. *Data Mining and Knowledge Discovery 24*(2), 355–386.

Schaller, N., E. Lazrak, P. Martin, J.-F. Mari, C. Aubry, et M. Benoît (2012). Combining farmers' decision rules and landscape stochastic regularities for landscape modelling. *Landscape Ecology 27*, 433–446.

Tatavarty, G., R. Bhatnagar, et B. Young (2007). Discovery of temporal dependencies between frequent patterns in multivariate time series. In *2007 IEEE Symposium on Computational Intelligence and Data Mining*, pp. 688–696.

Vlachos, M., P. Yu, et V. Castelli (2005). On periodicity detection and structural periodic similarity. In *Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005*.

Summary

This paper proposes a stochastic modeling of time series of items to mine periodic items into a stationary process. The representation of time series of items under the form of time series of tensors allows the use of digital signal processing methods like spectral analysis of auto cross correlation coefficients. On synthetic data, a spectral analysis allows the extraction of periodic behaviors.